DOCUMENT RESUME

ED 346 165 TM 018 649

AUTHOR Wise, Lauress L.
TITLE Lessons Learned from Military Performance Assessment.
INSTITUTION Defense Manpower Data Center, Monterey, CA.
PUB DATE 17 Apr 92
NOTE 17p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 21-23, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Armed Forces; Cost Effectiveness; *Educational Assessment; Federal Programs; Generalization; *Job Performance; Job Skills; Measurement Techniques; *Military Personnel; Military Training; *Occupational Tests; *Personnel Evaluation; Personnel Selection; Scaling; Scores; Test Construction; Test Reliability; Test Validity
IDENTIFIERS Hands On Performance Test (JPM Project); *Job Performance Measurement Project; *Performance Based Evaluation; Standard Setting

ABSTRACT

Lessons derived from the Job Performance Measurement (JPM) Project, which is overseen by the Office of the Assistant Secretary of Defense for Force Management and Personnel, for educational assessment are explored. The JPM Project was initiated to develop high fidelity measures of performance on the job that can be used to evaluate personnel selection and training procedures for the military. Several approaches to measuring performance of recruits were considered. The primary measure developed for each job was the hands-on performance test (HOPT) using a sample of critical tasks for each job. Each of the military services also developed several other types of measures. Reliability, validity, and cost issues are discussed for these measures and the HOPT. Ongoing issues include domains of interest, score scales, generalizing measures to other jobs, and setting performance standards. Among the many implications of the JPM Project for education are the following: (1) careful specification of the domain to be assessed is critical; (2) understanding the uses of the measures should precede their design; (3) attention should be paid to the type of scale required; (4) written tests are not enough; (5) alternative measures are expensive to develop, administer, and score; (6) more attention should be given to analogs of performance ratings; and (7) procedures for assessing the generalizability of performance measures are important. Two tables, 2 figures, and 16 references are included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED346165

LESSONS LEARNED FROM MILITARY PERFORMANCE ASSESSMENT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

LAURESS L. WISE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Dr. Lauress L. Wise
Defense Manpower Data Center

Paper Presented To:
National Council on Measurement in Education

21 April 1992
San Francisco, CA

Note: The views expressed are those of the authors and do not necessarily reflect the position of the Defense Manpower Data Center or other agencies within the Department of Defense.

TM018649

BEST COPY AVAILABLE

LESSONS LEARNED FROM MILITARY PERFORMANCE ASSESSMENT

Lauress L. Wise
Defense Manpower Data Center

Introduction

Last year, the U.S. military enlisted over 200,000 young men and women into active duty, provided them with occupational training, and put them to work as clerks, cooks, mechanics, medics, truck drivers, tank crew members and in well over five hundred other entry jobs. The total number of youth enlisted is down substantially from the level of 300,000 just a few years ago and it will surely go down further over the next several years. Nevertheless, the task of selecting such a large number of untrained youth and assigning them to so many different jobs is enormous.

In 1979, it was discovered that the Armed Services Vocational Aptitude Battery, the instrument used to select and assign enlistees, was misnormed. As a consequence, a large number of those who were enlisted were ill prepared to succeed in training and perform well on the job. A number of steps were taken in response to this problem. One of the most significant steps was the initiation of the Job Performance Measurement (JPM) Project. The purpose of this project was to develop high fidelity measures of performance on the job that could be used to evaluate selection and training procedures and to provide a basis for more precisely determining the level of cognitive and other skills required for success in different jobs.

The JPM Project is now over ten years old. A committee of the National Academy of Sciences (NAS), formed to provide technical oversight, recently issued a two volume book describing accomplishments to date (Wigdor & Green, Performance Assessment for the Workplace, 1991). Given the current intense and well-publicized interest in developing new and better educational achievement measures, it is reasonable to ask what lessons were learned in the JPM Project that may be applied to the assessment of educational achievement. This paper presents a brief review of the JPM Project, describes some of the issues that were encountered, and then discusses some of the lessons that might be applied in the educational arena.

Description of the JPM Project

The Job Performance Measurement Project was a coordinated effort involving each of the Armed Services, overseen by the Office of the Assistant Secretary of Defense for Force Management and Personnel (FM&P). Each Service selected a sample of jobs to represent the broad domain of occupational specialties available to their recruits. Each of the Services organizes its entry jobs somewhat differently, but four broad categories may be used to categorize jobs in terms of their cognitive requirements. These are: (1) Mechanical, (2)

Administrative, (3) Electrical/Electronic, and (4) General. Job classification systems used by the Air Force and Marine Corps use these four categories. The Army further divides the General into combat, technical, communications, operators, and general maintenance; the Navy also has further subdivisions within the mechanical and electrical families. The sample of jobs selected by the Services represented the different job categories used in each of their selection and classification systems. Table 1 lists the jobs included in the JPM project by Service and by job category.

The selection of jobs was based on a number of concerns including representation of the domain of ability requirements, emphasis on jobs with a large number of incumbents (both to assure adequate sample sizes and because the adequacy of selection procedures for such jobs affects more individuals), and emphasis on jobs judged to be most critical to the Service's overall mission. The Army conducted a larger effort, known as Project A, that included an expanded sample of jobs for which less extensive performance measures were developed.

Development of Job Performance Measures

A number of different approaches to measuring the performance of recruits in different jobs were considered. It is common to believe that measures of output provide the most direct assessment of the value of an individual's performance. Output measures were judged not to be appropriate for the JPM project primarily because: (1) level of output is influenced by a number of environmental and organizational factors unrelated to the individual (and the selection of individuals was the chief concern in this effort) and (2) many critical jobs do not have easily identifiable outputs, particularly combat specialties where readiness is all that can typically be measured.

The primary measure of performance developed for each job was the hands-on performance test (HOPT). A sample of critical tasks was identified for each job and incumbents were asked to perform each of these tasks. A scoring system was developed for each task that generally consisted of a number of separate steps on which the incumbent was scored as passing or failing (GO or NO-GO). Scorers were selected and trained to observe and score the performance on each of the selected tasks. This approach was selected as providing the most valid assessment of how well incumbents could perform their job (Wigdor & Green, 1991).

Each of the Services also developed a number of other types of measures. Some of these measures were evaluated as less expensive surrogates for the higher fidelity proficiency measure. Written job knowledge tests and walk-through performance simulations were examples of such measures. Other measures, particularly those developed by the Army, were designed to provide an assessment of a broader domain of performance. These included ratings of "typical" performance (as contrasted to proficiency tests that measured "maximal" performance) and indicators of achievement from administrative records. The primary comparison in the present discussion will be between the hands-on "work sample" measures, the written job knowledge measures, and the performance ratings. A brief description of the development of each of these measures is provided here. More detail may be found in Wigdor & Green (1991) and C. Campbell et al (1990).

Hands-On Performance Tests (HOPT). Extensive job analyses was performed for each of the jobs included in the JPM study resulting in lists of roughly 500 to 1,000 "tasks" identified in training materials, job-specific manuals, or occupational surveys. Tasks in the initial task lists varied considerably in detail and complexity. Simpler tasks were combined and more peripheral tasks were eliminated to narrow the initial list down to a revised list of from 150 to 250 tasks. Subject-matter-experts (SMEs) rated the criticality, difficulty, and frequency of each of the tasks and sorted them into separate content groups. A final sample of tasks was selected from the revised list for each job based on the information provided by the SMEs. The final number of tasks varied somewhat by Service. The Army selected 15 tasks for each job. The Marine Corps project included as many as 35 distinct tasks for one job. The final sample for each job was designed to emphasize the most critical and frequently performed tasks, to cover each of the different content areas identified in the task sorting exercises, and to avoid tasks that were so easy or difficult that little information on individual differences could be gained. For some jobs, an assessment of testing feasibility also was used in deciding among equally relevant tasks.

Once the tasks had been selected, scoring rules were developed for each task. Critical steps or behaviors were identified in consultation with SMEs. For the most part each step or behavior was scored as "GO" if it was performed successfully and "NO-GO" if it was not. In some cases, more quantitative information (e.g., typing speed, time to completion, number of targets hit) was also incorporated. In most cases, the total score for the task was the percent of steps/behaviors performed successfully. Figure 1 shows an example of a scoring sheet for one of the truck driver tasks.

In addition to developing the specific scoring rubric for each task, the project team developed a program for training scorers on the use of the scoring rubric. Generally tasks were organized into four to eight stations with a different scorer assigned to each station. Scorers received up to two days of training in scoring the tasks for the station(s) to which they were assigned. Special studies were conducted that included use of multiple scorers at each station and planned rotation of scorers through all stations.

Job Knowledge Tests. The written job knowledge tests were organized around the sample of tasks for which HOPT were developed. The purpose of this approach was to minimize measurement differences that might result from a different sampling of job content. In general, three to ten questions were written for each task. The questions tended to tap knowledge of correct procedures, correct sequencing, or standards to be met in performing the task and were not primarily concerned with task vocabulary. The Army also developed tests for 15 additional tasks for each job and a separate knowledge test organized around the training curriculum rather than the job analysis. Subsequent analyses indicated no significant differences between the original and supplemental tasks and between the task-based and training-based knowledge tests (McHenry, et al., 1990).

Rating Scales. An independent approach to job analyses was used in developing the performance rating scales. The critical incident technique (Flanagan, 1954), was used to elicit examples of specific behaviors judged by SMEs to be particularly effective or ineffective. Hundreds of incidents were collected for each job, edited into a common format, and clustered

into eight to fifteen categories. SMEs then participated in "retranslation" exercises where the edited incidents were matched to preliminary category descriptions and rated with regard to the level of effectiveness that they exemplified. Categories were combined or eliminated where there was some ambiguity about the matches of incidents to categories. Incidents for which there was high agreement as to the category and level of performance they exemplified were considered for use as scale anchors. A separate rating scale was constructed for each of the final incident categories. For each scale, one or more incidents were summarized to describe particularly effective performance (scale levels 6 and 7), particularly ineffective performance (scale levels 1 and 2), and average performance (levels 3, 4, and 5). In some cases there were not many incidents at the middle levels of performance and the behavioral summaries for the middle performance level were synthesized to be consistent with the summaries at either end of the scale. Figure 2 provides an example of one of the rating scales that resulted from the process.

In addition to developing well-anchored behavioral scales, the Project A staff also developed a rater training program (Pulakos & Borman, 1986). This training program was designed to reduce halo effects and to promote more accurate comparisons among those rated, building on the results of previous research on similar programs (Borman, 1979).

Other Measures. The Air Force developed a set of "walk through" performance simulations for tasks were actual performance would be too costly, time-consuming, or dangerous. In these simulations, the examinee describes how the task would be performed. These descriptions are the rated for accuracy and completeness as if the examinee had actually performed the task.

The Army conducted an extensive investigation of the use of administrative measures including awards and certificates, letters of commendation, promotion rates, and disciplinary actions. These measures were found to agree more closely with ratings than with the hands-on and written proficiency measures (J. Campbell, et. al, 1990).

In extending its performance measurement efforts to cover soldiers in their second tour of duty, the Army also is developing measures of supervisory performance that include role-play simulations. These measures involve ratings of live interactions with confederates who are to be counseled or disciplined. While preliminary tryouts of these measures show some promise, final data are still being collected.

Results

Gottfredson (1990) reviewed five dimensions that may be used in evaluating criterion measures. These include: validity, reliability, susceptibility to compromise, cost, and acceptability to interested parties. How did the different measures developed in the JPM efforts fare against these criteria? A brief discussion of the evaluation of the three main types of measures, HOPT, ratings, and written tests, against these dimension is presented below, followed by an overview of on-going issues in the development and use of military job performance measures.

Validity. Validity applies to the use of a measure and not just to the measure itself. Measures that are used to predict future outcomes can be validated by demonstrating the statistical relationship between the measure and the outcomes it is used to predict. With criterion measures, validation is necessarily more judgmental and typically involves a combination of argument concerning the relationship of the content of the measure and the domain of interest and analyses of the similarities and differences among different measures of relevance.

The hands-on performance tests (HOPT) are essentially work samples and, as such, are necessarily assessments of the construct of interest. The validity of the HOPT rests primarily on arguments concerning the representitiveness of the tasks that were sampled. This is particularly true because the results of generalizability analyses (discussed under below) indicated a significant interaction between examinees and tasks. The domain of task performance is thus not clearly unidimensional, with each task providing an equivalent ordering of examinees, and so the representativeness of the tasks that were sampled is at issue. Fortunately, the careful task sampling procedures used in the JPM project provide a high degree of support for the validity of the measures that resulted.

The performance rating measures also have a high degree of validity based on the systematic procedures used in their development and on SME evaluations of the adequacy of their coverage of the job content. Examinees are ordered somewhat differently by the ratings in comparison to the hands-on performance tests (J.Campbell, et al, 1990). The relative validity of the two measures depends upon the exact construct of interest. If one is interested in how well incumbents do actually perform on the job and not just how well they could perform, then a good argument could be made for the higher validity of performance ratings since they assess routine performance and not just performance on one special occasion.

The validity of written measures is somewhat more questionable on two grounds. First, for many tasks there are obvious differences between knowing how to perform the task and being able to actually perform it successfully. Written tests may thus be viewed as incomplete measures of the domain of interest. The second concern with the written measures is that they include irrelevant sources of variation. In particular, measures of verbal ability used in selection were much more highly correlated with the written performance test scores than they were with the HOPTs. Nonetheless, the correlations of the written measures with the HOPT were quite high for most jobs, suggesting that such differences may not have large practical significance.

Reliability. The reliability of all of the measures described here is quite high. Table 2 shows the range of reliabilities reported for the HOPT. It should be noted that these reliabilities are for mean scores across a relatively large number (10 to 35) of different tasks. A detailed analysis of components of error was conducted for two jobs (Webb, et al, 1989; Shavelson, et al, 1990). The primary issue was the level of agreement among scorers. The results showed virtually no contribution to error from scorer differences, indicating success in training the scorers and in creating objectively scorable measures. The results did indicate, however, that there were very significant mean differences among different tasks and even more significant task by examinee interactions. For riflemen, the variance component

estimates were .011 for examinee, .034 for task, .070 for examinee by task, and .012 for error leading to relative and absolute G coefficients of .12 and .08 for performance on a single task. With a relatively large number of different tasks included, the relative and absolute G coefficients were as high as .83 and .77 for riflemen and .72 and .65 for machinist mates (Shavelson, et al, 1990).

Reliability estimates for the performance ratings were acceptably high, in the range of .86 to .93 (Campbell, et al, 1986; Pulakos & Borman, 1986). These high reliability estimates were obtained for averages across several different raters. The reliability for individual raters was significantly lower.

The written performance tests had the highest levels of estimated reliability, generally in the range of .75 to .91. These estimates were generally based on internal consistency analyses (coefficient alpha). If the domain measured is significantly multi-dimensional, parallel forms reliabilities could be even higher. One short-coming of the JPM research is that detailed reliability estimates were available for only some of the measures some of the time. It would be useful to have more complete analyses of sources of variation/generalizability for all of the measures and to include variation across occasions (test-retest) as a source of error.

Susceptibility to Compromise. The JPM effort did not involve "operational" use of the job performance measures so motivation for compromise was limited. Rating measures that are in use for operational performance assessment show very significant skew and ceiling effects that are much different from the JPM results. It is possible that rater training procedures might protect against problems with operational use of ratings.

It would be difficult to "cheat" on the hands-on measures, in that an answer key would be of minimal use. One problem with the hands-on tests was that they were necessarily administered "one-on-one". Consequently, not everyone was tested at the same time and, in a few instances, there were issues as to whether examinees had an opportunity to observe others performing the tasks prior to their being tested and whether some examinees had sufficient prior knowledge of the tasks to be tested (testing extended over several days) and were able to practice these tasks in advance. If such selective practice did occur, generalization from the specific tasks tested to the entire domain would be jeopardized.

Written tests have long been used operationally, in part, because they can be relatively secure from compromise. Examinees can be tested all at once protecting against prior exposure and the tests include a large number of items so that selective practice is not generally feasible.

Cost. HOPT were costly to develop and very costly to administer. Administration was one-on-one (and sometimes two-on-one) so the costs of the scorers was enormous. Examinee time also was an issue as four to eight hours were required to complete the testing for most jobs. Equipment and testing space costs also were significant for most jobs. Exact costs associated with this mode of testing are difficult to pin down, but were in the millions of dollars (without counting examinee time) even for the relatively limited samples of jobs and examinees tested in the JPM project.

By comparison, ratings were much less expensive to administer. The ratings did, however, involve assembling and training multiple supervisors for each examinee. Supervisors are generally a scarcer resource in comparison to the job incumbents. Frequently, however, one supervisor could rate several different subordinates so that the total number of supervisors was generally less than the number of examinees. Rating sessions generally lasted one to two hours.

We knew, going in, that written tests were the least expensive to develop and administer. There were no surprises in this regard. The tests were administered in groups of up to 50 or more, so administrator costs were minimal. In addition, the test required only an hour or two of examinee time compared to four to eight hours for the hands-on measures.

Acceptability to Interested Parties. There was not a systematic attempt to collect information on the acceptability of the different measures to military personnel managers, however, the Army effort did include some assessment of the credibility of different measures during their tryout (Campbell, 1987). The HOPT and supervisor ratings both had very high levels of acceptability to military personnel managers. Peer ratings were somewhat less acceptable, notwithstanding empirical evidence that they gave the same results as the supervisor ratings. Evaluations of the written measures were mixed.

On-Going Issues

Given the number of groups and organizations that participated in the JPM Project, complete agreement on either the approach to be taken or on the interpretation of the results would be surprising, and there were no such surprises. The present paper reflects the views and interpretations of the author and is not necessarily endorsed by any of the Services, the NAS Committee that worked with the Services, or by the Office of Secretary of Defense which sponsored and worked with the Committee and the Services. A brief discussion of some of major areas of contention is presented here before turning to an analysis of lessons learned that may be relevant to educational measurement.

Domain of Interest. Their was considerable debate and discussion regarding the definition of the domain of performance that was of interest. The debate ranged from a relatively narrow focus on job "proficiency" to much more expansive conceptions of "productivity". The NAS Committee attempted to focus the project on a concept of job proficiency. Selection procedures and standards could be expected to predict an individual's proficiency; other aspects of productivity depended on organizational and environmental conditions not under the individual's control and therefore not relevant to the procedures and standards used in selecting individuals. The Army took a different perspective believing that productivity was a function of effort and discipline as well as proficiency.

The staff working on Project A developed a more elaborate model of job performance that subsumed proficiency as one component (J. Campbell, et al, 1990). The HOPT and JKT measures were found to be appropriate measures of "proficiency", assessing what the individual could do under optimal conditions. The rating and administrative measures, on the other hand, assessed "typical" performance that reflected a combination of proficiency, effort,

and discipline. One consequence of this different perspective was the finding that temperament measures such as achievement orientation and dependability did not add significantly to the prediction of proficiency but added very significantly to the prediction of other aspects of performance (e.g., effort and discipline) (McHenry, et al, 1990).

Score Scale. A second area of some debate was the use to be made of the score scale. One point of view was that the primary use of the performance scores was in validating selection procedures. The performance measures should measure individual differences as accurately as possible so that the extent to which these differences were related to the selection tests could be determined. For this purpose only a relative scale was required. Items or tasks that were too easy or too hard were not judged useful since they provided little information on individual differences. The NAS Committee, on the other hand, argued for more extensive considerations that included setting standards for performance and comparing them across jobs. To this end a more absolute scale, assessing the percent of the job the incumbent could successfully perform, was desired. The NAS Committee argued for a random sample of tasks, possibly stratified, from the entire domain of performance. Easy or hard tasks should not be excluded as this would bias the level of the resulting scale.

One area of general agreement concerning the score scale was that it should be finitely bounded at both ends, ranging from a zero point (or lower bound to the rating scales) to a fixed maximum (100 percent or an upper bound for the rating scales). Those desiring an emphasis on individual differences did not argue for an IRT "theta" scale where performance levels stretch out infinitely in either direction. The performance domain, whether broadly or narrowly defined, was assumed to be multidimensional and the steps within each task had obvious dependencies so that IRT approaches were not attempted.

Generalizing to Other Jobs. Data from Project A were used to demonstrate differences in predictive relationships across jobs (Wise, et al., 1990). Whetzel (1991) demonstrated that such differences in prediction can have a significant impact on classification efficiency. Alternative procedures for extending the results from the JPM job sample to all entry-level military specialties have been explored. The second phase of the DoD-wide JPM effort, known as the "Linkage Project", used an approach that involved: (1) use of an occupational "cross-walk" that matched military occupations to civilian occupations; (2) development of job complexity factors from information available in the Dictionary of Occupational Titles and estimation of factor scores for the civilian counterparts to each military occupation; (3) fitting a hierarchical linear model that relates both individual characteristics (test scores, education level, and job experience) and job characteristics (complexity factor scores) to job performance levels using data on the JPM job samples (Harris, et al., 1991).

In a separate effort, the Army Synthetic Validation Project developed prediction equations (predictor composites) for each of 96 task categories and designed a rating system for developing job-specific category weights for use in developing a composite prediction equation. A "job-family" approach also was explored as an alternative to developing separate prediction equations for every job.

Neither of these approaches has yet been used operationally in developing job-specific selection composites. Each of the Services is continuing to explore revision of their selection composites based different approaches to generalizing predictive relationships and to identifying job families.

Setting Performance Standards. The two major follow-on projects also made attempts to explore the use of the new performance measures for setting performance standards. The Joint-Service Linkage Project involved efforts to put the HOPT percent-Go scores onto a common scale and to generalize the relationship of predictor scores to these performance measures across jobs. This effort has not, however, set explicit standards for job performance, but has sought to understand an optimum mix of performance levels through cost-benefit analyses that assume a linear relationship between performance and benefit (Harris, et al, 1991).

In the Army Synthetic Validation Project, performance standards were set using the HOPT and also using the incident descriptions generated in the development of the rating scales. A major challenge was the use of measures developed for one job in setting performance standards for other jobs. Success in this endeavor was limited at best. One finding was that similar standards were set for all jobs with between 60% and 70% correct defining the minimum for acceptable performance. Such standards may well generalize if the measures are appropriately referenced to the underlying domain of performance and not subject to more arbitrary variation in difficulty levels. The critical incident approach and the performance categories that they represented were also moderately successful in providing a basis for generalized performance standards (Wise, et al., 1991).

Lessons for Educational Measurement

While many issues continue to be debated, it is clear that the Job Performance Measurement Project was quite successful in most areas. The performance measures that were developed had high validity and were acceptably reliable. Much was learned from similarities and differences among different types of measures including a better understanding of what it means to perform well in a job. Current selection tests were shown to predict the "authentic" measures of performance well, although not quite as well as they predicted paper-and-pencil job knowledge test scores. Room was left for a greater role for measures of achievement orientation and dependability and perhaps for spatial and psychomotor skills in the selection process. The jury is still out on the question of setting performance and selection standards.

What do the results of the JPM project have to say to those seeking to develop more authentic measures of educational achievement? Among the many implications of the JPM results are the following:

1. A careful specification of the domain to be assessed is critical. In the military arena there was a long history of job analysis that provided a good basis for defining the measurement domain. The biggest debates were over whether measurement should reflect

maximal performance (labelled proficiency by the NAS Committee) or typical performance and be broadened to include aspects of effectiveness such as teamwork or performance under adverse conditions. Results showing significant examinee-by-task interactions also suggest the need to assure appropriate coverage of the target domain.

In the educational arena, the relative success in defining the domain of mathematics for the state-by-state comparisons was surprising to many. Building a consensus for other subjects may be somewhat more troublesome. The debate about whether and how to measure typical performance rather than just maximal performance is also very relevant. There is an explicit desire to create measures that reflect and hence motivate effort -- an exam the student can and should study for.

2. An understanding of the uses to be made of the measures should precede their design. The JPM project experienced tension between those who wanted to use the scores primarily to validate selection tests and those who also wanted to set performance standards that might be generalized across jobs. The former group was content with maximizing the reliable variability in individual differences. All they wanted was a good norm-referenced test. The latter group, wanted a carefully constructed, domain-referenced measure. The former group wanted an interval level scale for their correlational analyses. The latter group might have been satisfied with an ordinal scale for comparing individuals to standards, but actually argued for an absolute scale in order to simplify generalizations across jobs.

Multiple uses also have been discussed for improved educational achievement measures. If our purpose is to evaluate and improve educational systems, norm-referenced scales should suffice. If we want to address questions like how much education is enough (or how much should we spend on education), for either individuals or for the system as a whole, criterion-referenced scales may be more appropriate.

3. Attention should be paid to the type of scale that is required. The JPM performance scales had multiple uses. For some uses, such as dividing performance into acceptable versus unacceptable, an ordinal scale would suffice. For other purposes, such as placing dollar values on the utility of different performance levels, a ratio scale would be required. Item response theory (IRT) was developed in the educational arena. It provides a basis for generalizing scores across different sets of items or different samples of examinees. Many believe that it provides an "interval" scale of achievement. This is only technically true, within the framework of the measurement model. A different measurement model with monotonic transformations of the theta scale (and the associated item characteristic curves) could fit the data equally well. IRT models may not work well with more complex behavioral samples involving greater interdependencies among the scorable units. Also, a scale that suggests an "infinite" difference between knowing nothing and knowing a little may not be appropriate. The percent-GO, or percent-correct metric used in the JPM project provides a reasonable scale for competency assessment and should be considered in more complex educational assessment as well.

4. Written tests are not everything. Knowing facts about a procedure and being able to execute it successfully are not exactly the same thing. The results varied considerably

across and within jobs, but generally sufficient differences were found between written and hands-on performance tests to encourage educational researchers to press on in efforts to develop alternatives. So far, however, very little has been changed as a result of developing and using alternative measures.

5. Alternatives measures are expensive to develop and even more expensive to administer and score. This finding is not at all new to those who have been working on standardized writing assessments. Difficulties in developing equivalent prompts, in hiring and training scorers and the great time and costs associated with reading and scoring large numbers of essays are well known. In a time of great competition for educational dollars, the potentially modest benefits associated with better measures must be carefully weighed against the costs.

6. More attention might be given to analogs of performance ratings. Tests are one-time events. A "portfolio" approach offers some promise for assessing typical rather than just maximal performance, but well-developed rating scales and well-designed rater training programs may offer a lower cost alternative. Grades might be viewed as a form of rating, but frequently the desire for "objectivity" has removed any component of expert judgment. What is lacking is the type of standardization across teachers and schools that might be gained through carefully constructed and anchored scales. Other principles, such as the use of multiple raters and of rater training programs might also be integrated.

7. Procedures for assessing the generalizability of performance measures are important. Results from the JPM effort indicated very significant examinee-by-task interactions. Military personnel researchers are still debating the extent to which performance measures can be generalized across jobs. This also is a key issue in educational assessment. More novel performance tasks may not generalize well outside a specific domain of knowledge.

The above list merely scratches the surface of lessons that the military job performance work might have for those seeking to develop better measures of educational achievement. Interested researchers should review of the two-volume report of the NAS Committee available from the National Academy Press and the 1990 issue of Personnel Psychology devoted to the Army's Project A as well as other articles and reports coming from this effort. Further information on this effort can be obtained from Dr. Jane Arabian, Assistant Director, Office of Accession Policy, OASD(fm&P), 2B271, The Pentagon.

References

Borman, W.C. (1979). Format and training effects on rating accuracy and rating errors. Journal of Applied Psychology, 64, 412-421.

Campbell, C.H., Campbell, R.C., Rumsey, M.G. & Edwards, D.C. (1986). Development and field test of task-based MOS-specific criterion measures (ARI Technical Report 717). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Campbell, C.H., Ford, P. Rumsey, M.G., Pulakos, E.D., Borman, W. C., Felker, D. B., de Vera, M. V., & Reigelhaupt, B. J. (1990). Development of multiple job performance measures in a representative sample of Jobs. Personnel Psychology, 43, 277-300.

Campbell, J.P. (Ed.). (1987). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1985 fiscal year. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Campbell, J.P., McHenry, J.J., & Wise, L. L. (1990). Modeling job performance in a population of jobs. Personnel Psychology, 43, 313-334.

Flanagan, J.C. (1954). The critical incident technique. Psychological Bulletin, 51, 327-358.

Gottfredson, L. (1990). The evaluation of alternative measures of job performance. In A.K. Wigdor & B.F. Green (Eds.), Performance assessment for the workplace, Vol. 2. Washington, DC: National Academy Press.

Harris, D.A., McCloy, R.A., Dempsey, J.R., Roth, C., Sackett, P.R., Hedges, L.V., Smith, D.A., & Hogan, P.F. (1991). Linking enlistment standards to job performance, Phase I: A job performance model. Alexandria, VA: Human Resources Research Organization.

McHenry, J.J., Hough, L.M., Toquan, J.L, Hanson, M.A. & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. Personnel Psychology, 43, 335-354.

Pulakos, E.D. & Borman, W.C. (1986). Development and field test of Army-wide rating scales and rater orientation and training program (ARI Technical Report 716). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Shavelson, R.J., Mayberry, P.W., Weichang, L. & Webb, N.M. (1990). Generalizability of job performance measurements: Marine Corps rifleman. Military Psychology, 2, 129-144.

Web, N.M., Shavelson, R.J., Kim, K.-S, & Chen, Z. (1989). Reliability (generalizability) of job performance measurements: Navy machinist mates. Military Psychology, 1, 91-110.

Whetzel, D. L. (1991). Multidimensional screening: Comparison of a single-stage personnel selection/classification process with alternative strategies. Unpublished Doctoral Dissertation, George Washington University, Washington DC.

Wigdor, A. K. & Green, B. F. (Eds.) (1991). Performance Assessment for the Workplace. Washington, DC: National Academy Press.

Wise, L.L., McHenry, J.J., & Campbell, J.P. (1990). Identifying optimal predictor composites and testing for generalizability across jobs and performance factors. Personnel Psychology, 43, 355-366.

Wise, L.L, Peterson, N.G., Hoffman, R.G., Campbell, J.P., & Arabian, J.M. (1991). Army Synthetic Validity Project: Report of Phase III results. (ARI Technical Report 922). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Table 1. Occupational Specialties Included in the Job Performance Measurement Project

Service	Job Family	Occupational Title
Army	General	Infantryman
	General	Cannon Crewman
	General	Tank Crewman
	General	Radio Teletype Operator
	Administrative	Medical Specialist
	Mechanical	Light Wheel Vehicle/Power Generator Mechanic
	General	Motor Transport Operator
	Administrative	Administrative Specialist
	General	Military Police
Navy	Mechanical	Machinists Mate
	General	Radioman
	Electrical	Electronics Technician
	Electrical	Electrician's Mate
	General	Fire Control Specialist
	Mechanical	Gas Turbine Technician, Mechanical
Air Force	Mechanical	Jet Engine Mechanic
	Mechanical	Aerospace Ground Equipment Mechanic
	Administrative	Personnel Specialist
	Administrative	Information Systems Radio Operator
	General	Air Traffic Control Operator
	General	Aircrew Life Support Specialist
	Electrical	Precision Measurement Laboratory Equipment Specialist
	Electrical	Avionic Communications Specialist
Marine Corps	General	Infantry (Rifleman, Machinegunner, Mortarman, Assaultman, Unit Leader)
	Mechanical	Automotive Mechanic
	Mechanical	Helicopter Mechanic

Table 2. Mean/Median Reliability Estimates for Each Measure

Performance Measure	Reliability Estimates
Hands-On Performance Tests*	.75 to .94 (mean .82)
Written Job Knowledge Tests**	.85
Behaviorally Anchored Rating Scales**	.59 (Single Rater) .89 (Mean of 6 Raters)

Source:
* - Wigdor & Green (1991, p. 121)
** - C. Campbell, et al. (1990).

Instructions to Examinee

"During this test you must drive the tractor and semitrailer through the course. (Explain layout of course.) You do not have to perform PMCS [preventative maintenance checks]. You must perform each maneuver without assistance from me or the scorer/ assistant. You may adjust the seat and mirrors before we begin the test. Tell me when you are ready.

Operate Tractor and Semitrailer Task Shoresheet

	Steps	GO	NO-GO
1.	Selected first gear position.	____	____
2.	Maintained RPM (did not allow vehicle to lug, jerk, or stall).	____	____
3.	Shifted gears without grinding. (Mark N/A for M915)	____	____
4.	Drove without riding clutch. (Mark N/A for M915)	____	____
5.	Kept both hands on the wheel (except when shifting.	____	____
6.	Used hand over hand steering when turning.	____	____
7.	Passed restricted roadway without striking roadway markers.	____	____
8.	Followed serpentine roadway.	____	____
9.	Passed serpentine without striking barriers.	____	____

NOTE TO SCORER: Do not allow soldier to stop before entering the restricted roadway and serpentine. Soldier should maintain 10-15 MPH going through both obstacles.

Figure 1. Example of Hands-On Performance Test Score Sheet

D. Using Maps/Following Proper Routes

How effective is each soldier in securing proper maps as needed; becoming familiar with routes ahead of time when appropriate; using maps effectively; following prescribed routes; and arriving at commitments on time?

Sometimes fails to secure or use maps when needed; may not be able to read maps properly; is often late i n reaching designated location. Sometimes fail to plan route ahead of time; may take unplanned route.	Almost always secures maps if needed and uses most maps effectively; usually completes commitments on time. Almost always plans route ahead of time; generally becomes familiar with route before commitment; usually follows planned route.	Always obtains proper maps when needed; is able to use grid coordinates to reach even hard-to-find sites; always completes commitments on time. Always plans and becomes familiar with route before commitments; always follows planned route.
1 2	3 4 5	6 7

Figure 2. Example of Behaviorally Anchored Rating Scale

17 BEST COPY AVAILABLE